Indexing Key Positions between Multiple Videos

Navneet Dalal, Radu Horaud

Indexing Key Positions between Multiple Videos

Navneet Dalal Radu Horaud

INRIA Rhône-Alpes
655 avenue de l'Europe-38334, Montbonnot, FRANCE
E-mail: Navneet.Dalal, Radu.Horaud@inrialpes.fr

Abstract

Given two or more video sequences containing similar human activities (running, jumping, etc.) we want to devise a method which extracts spatio-temporal signatures associated with these activities, compares these signatures, and aligns key positions of different videos. In this paper we introduce a method which, in conjunction with a number of hypotheses, allows the analysis of the motion of specific body parts and extracts their 2D (image plane) time-varying trajectories as well as their 3D trajectories. Two such trajectories recovered from two different videos have different characteristics. We develop a curve registration technique which consists of estimating a transformation mapping one time-basis (of the first curve) onto another time-basis (the second curve). We also analyse in depth the conditions under which such curve registration techniques are valid. Finally, we show results with two similar athletic events performed by two different athletes.

1 Introduction

In this paper we address the following problem. Given two or more video sequences containing similar human activities (walking, running, jumping, etc.) we want to devise a method which extracts spatio-temporal signatures associated with these activities, compares these signatures, and aligns key positions of different videos. An example domain of such an application is sport events. Currently, individual performances such as jumps and vaults (athletics, gymnastics, etc.) are difficult to compare quantitatively, from one athlete to another, because the only available data are video sequences.

Because we want to analyse activities with high dynamics occurring over several seconds, the camera must move and its settings (zoom and focus) must vary such that the human subject remains within the field of view. Therefore, the apparent image motion is a combination of camera motion (referred herein as egomotion) and 3D human-body motion. Full recovery of the 3D trajectory, kinematics and dynamics of the human body from a single moving camera remains an ambitious goal.

In this paper we introduce a method which, in conjunction with a number of hypotheses, allows the analysis of the motion of specific body parts (torso, hips, and so forth), and extracts their 2D (image plane) time-varying trajectories as well as their 3D trajectories. Two such trajectories recovered from two different videos have different characteristics. We develop a curve registration technique which consists of estimating a time warp function, i.e., a transformation mapping one time-basis (of the first curve) onto another time-basis (the second curve). Also, we analyse in depth the conditions under which this curve registration method returns valid results.

Without loss of generality we introduce a number of constraints. First, we restrict camera motion to pan and tilt rotations around the optical axis as well as variations in zoom and focus. This type of camera motion is conveniently described by a plane-to-plane projective transformation and can be estimated from image point correspondences without any prior knowledge. Cameras mounted on tripod (as used in athletic events) satisfy these conditons to a good accuracy [1]. Second, we assume that image regions corresponding to individual body parts can be tracked through the video sequence. Third, we assume that the focal length varies in a way such that the apparent image size of the object of interest (here a human body) remains constant. The latter implies that variations in focal length compensate for variations in depth. Interesting enough, not only that this hypothesis allows the recovery of the full 3D trajectory of a body part, but it is verified in practice.

The concept of mapping one time-basis onto another time-basis (time warping) is a well known technique in the engineering literature. Initial applications to speech-processing can be found in [7]. Applications in computer vision to human gesture recognition can be found in [2, 3]. In [3] authors used learned view-dependent models and performed time warping using pattern matching techniques known in speech recognition. Hidden Markov Models (HMMs) are also considered for action recognition. In [12], HMMs have been used to recognize American Sign Language from videos. The advantages of HMMs over

time warping is that HMMs can learn from the training data. However HMMs required network topology to be described and fine-tuned precisely. Whereas compared to HMMs, time warping is conceptually simpler and elegant. Still, vision community has not paid much attention to such works. One reason is viewpoint dependency of motion trajectories. In this work, we will show for the first time the conditions required for motion trajectories to be viewpoint invariant and validity of these conditions in real life video sequences. In [10], authors showed affine invariance of periodic motion. In comparison, we consider general motion and invariance conditions are shown for general perspective transformation. Second reason is reliable estimation of time warp functions. In earlier works, pattern recognition framework or landmark registration was used to estimate time warp functions. These registration techniques were not robust enough to do reliable registration [8, 9]. However, literature in statistics has developed much over these years with robust registration techniques like Continuous Curve Registration [8, 9] being applied to many different fields [9].

Also in earlier works, authors [2, 3] concentrated on gesture recognition, in this paper, we propose to use this framework for indexing key positions. We further believe that it can be easily extended to other vision applications like event recognition, human body joint prediction, human activity performance analysis, and comparison of movements.

The remainder of this paper is organized as follows. In §2 we briefly recall the method used to estimate the camera motion parameters from static scene points and the characteristics of the image-region (corresponding to human body parts) tracker being used. Next we describe how the initial trajectory associated with the moving image-region is transformed into an egomotion-compensated trajectory using adaptive manifold. In §3 we formulate our problem as curve registration w.r.t. each other such that when registered, at any time, curves correspond to identical state of the human activity. In §4 we look at sufficient and necessary conditions required for registration, their validity in real life video sequences. In §5, we present results of automatic indexing of key positions in video sequences. Finally, we finish with conclusions and future work in §6.

2 Extracting Curves from Image Sequences

We first show how to estimate image plane trajectories or curves needed for curve registration from video sequences. This requires: i) estimating camera motion; ii) tracking image-region (for e.g. athlete's torso); and iii) estimating its egomotion compensated image plane trajectory for whole video sequence.

In the rest of the paper, we use $\mathbf{x}(t) = [x(t), y(t)]^\mathsf{T}$ as the 2D centre position of the image-region at any time t and $\mathbf{y}(t)$ be the same position but egomotion compensated. For discrete case scenario, we use $\mathbf{x}^j = [x^j, y^j]^\mathsf{T}$ and $\mathbf{y}^j$ respectively. Here j denotes respective positions in j-th frame.

2.1 Camera Motion Estimation

We assume that camera's centre of projection is fixed and only focal length f, tilt angle θ_1 and pan angle θ_2 vary. These assumptions are valid for cameras mounted on tripods and hence sufficient for most video sequences [1].

Let $\mathsf{H}^{j(j-1)}$ be the plane-plane projective transformation or homography warping the $(j-1)$-th frame to current j-th frame. At each time instant j, we estimate homographies $\mathsf{H}^{j(j-1)}$ parameterized over f, θ_1 and θ_2 by minimizing dense photometric information between two consecutive frames [1, 11] and self-calibrate the sequence [4, 5] to obtain f, θ_1, θ_2 values for complete video.

2.2 Tracking

To estimate $\mathbf{x}^j$, we automatically track image-regions (corresponding to specific body parts for e.g. athlete's torso) in whole video sequence using CONDENSATION [6] based framework. Color histograms of image-regions are used as image based measurements during reactive reinforcement step. In the first frame, we manually initialize our tracker by selecting an image-region. In each subsequent frame, we take first order moment of the posteriori probability density as value of $\mathbf{x}^j$.

2.3 Using Adaptive Manifold

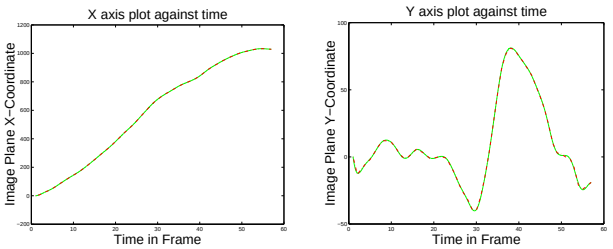
We now estimate egomotion compensated image plane trajectory $\mathbf{y}^j$ as spanned by an image-region. One possible solution can be to warp $\mathbf{x}^j$ in input images to a static manifold using the homographies obtained in §2.1. But note that if the static manifold is a plane (i.e. planar mosaic), then due to characteristic bow-tie shape of the mosaic, $\mathbf{x}^j$ far away from centre of planar mosaic will tend to blow up. And if one uses a cylindrical or spherical manifold, then x- and y- coordinates of $\mathbf{x}^j$ couple nonlinearly. Thus, the result depends on which manifold is used and which input image is used as a reference frame for building panoramas.

So we propose to subtract egomotion from input images on frame-by-frame basis. More formally, given the position $\mathbf{x}^j$ in j-th frame and $\mathbf{x}^{j-1}$ in $(j-1)$-th frame, its velocity $\mathbf{v}^j$ in j-th frame is

$$\mathbf{v}^j = \mathbf{x}^j - \Psi\left(\mathsf{H}^{j(j-1)} \begin{bmatrix} \mathbf{x}^{j-1} \\ 1 \end{bmatrix}\right) \qquad (1)$$

where Ψ converts homogenous point coordinates into Euclidian ones and the homography $\mathsf{H}^{j(j-1)}$ compensates for camera motion and zoom variations. Note that egomotion does not contribute to the term $\mathbf{v}^j$. If Δt is the time between two frames, then as

$$\lim_{\Delta t \to 0} \mathbf{v}^j = \mathbf{v}(t)$$



(a) (b)

Figure 1. Trajectory of torso of an athlete doing high jump. Curve $\mathbf{y}(t)$ obtained as described in §2.3: (a) x-coordinate of $\mathbf{y}(t)$ and (b) y-coordinate of $\mathbf{y}(t)$ plotted against time.

where $\mathbf{v}(t)$ is the true velocity of the tracked image-region in image plane. An integral of $\mathbf{v}(t)$, or summation of $\mathbf{v}^j$ in discrete case, will give us the required signature curve over time i.e.

$$\mathbf{y}^j = \sum_j \mathbf{v}^j \tag{2}$$

where $\mathbf{y}^j$ is the instantaneous path followed by the tracked image-region in the image plane. Curve registration techniques can now be applied to $\mathbf{y}^j$. Note that as $\mathbf{y}^j$ is evaluated frame to frame, unlike static manifold mosaic case, artifacts due to cumulative errors do not appear (as curvature of $\mathbf{y}(t)$ at any time depends only upon instantaneous time t).

From practical point of view, we would like to estimate the value of $\mathbf{y}^j$ at any time instance and also possibly its higher order derivatives. Therefore we fit a B-spline curve of order 6 to be able to estimate $\mathbf{y}^j$ up to its second order derivatives. Because of B-spline curve fitting, from now on, we will treat $\mathbf{y}^j$ as a continuous function of time i.e. $\mathbf{y}(t)$. Figure 1 shows one such $\mathbf{y}(t)$ for a high-jump sequence.

Note that $\mathbf{v}(t)$ is the perceived instantenous velocity in the image plane. Even though $\mathbf{y}(t)$ is free of the problems listed above, it does depend upon camera's viewpoint relative to original trajectory of the point in 3D space. In §4 we discuss conditions for viewpoint invariance.

3 Problem Formulation

Consider signature curves $\mathbf{y}_i(t_i), i \in [1, \ldots, N]$ corresponding to N similar events gathered from N different videos (high-jumps, pole-vaults, and so forth performed by different athletes at different times with different camera settings and viewpoint). The 2D vectors $\mathbf{y}_i(t_i)$ are as obtained in §2.3.

Let these functions be defined on closed real intervals that can be taken without loss of generality as $[0, T_i]$. The values $\mathbf{y}_i(t_i)$ of two or more functions may differ because of two types of variation: i) *amplitude variation* due to the fact that two functions $\mathbf{y}_1$ and $\mathbf{y}_2$ may simply differ at points of time at which they can be compared; ii) *phase variation* in the sense that $\mathbf{y}_1$ and $\mathbf{y}_2$ should not be compared at a fixed time t, but at times t_1 and t_2 at which the two events are essentially in physically comparable states, so that the curves exhibit comparable features at these times.

Let time interval $[0, T_0]$ be a standard or reference interval. Let $h_i(t)$ be a transformation of time t (or time warp function) for case i with domain $[0, T_0]$. The fact that the timings of events have the same order regardless of the time scale implies that h_i, the time warping function should be *strictly increasing* and hence *invertible*. Thus one can always solve the equation $t_i = h_i(t)$ for t given value t_i or vice-versa. Also, $h_i(t)$ must satisfy the boundary conditions $h_i(0) = 0$ and $h_i(T_0) = T_i$. In addition, one may require that $h_i(t)$ be a smooth function of t in the sense of being differentiable a certain number of times.

Let $\mathbf{y}_0(t), t \in [0, T_0]$ be a fixed function that provides a template or reference for the individual curves $\mathbf{y}_i$, i.e. after registration, the features of $\mathbf{y}_i$ will be aligned in some sense to those of $\mathbf{y}_0$. One may thus propose the following general model

$$\mathbf{y}_i[h_i(t)] = \mathbf{A}_i(t)\mathbf{y}_0(t) + \epsilon_i(t) \text{ or } \mathbf{y}_i \circ h_i = \mathbf{A}_i\mathbf{y}_0 + \epsilon_i, \tag{3}$$

where ϵ is 2D error (relatively small as compared to $\mathbf{y}_i$ and roughly centred about 0), and $\mathbf{A}_i$ is 2D amplitude modulation function.

The registration task, then, is to estimate the time warping functions h_i so that the de-warped components $\mathbf{y}_i$ can be studied separately, along with possible analysis of the functions h_i as well. We estimate the time warp functions using continuous curve registration techniques [8] as described below.

3.1 Continuous Curve Registration

We describe continuous curve registration using scalar functions (see [8]). For vector valued functions, one can form a composite criterion by adding registration criteria function (5) across all dimensions.

In the following text, we will drop the subscript i for simplicity of notation. Now consider curve values as a set of points to which one might apply principal components analysis. If curve values are proportional, then such an analysis would yield only one nonzero principal component. That is, only one of the two eigenvalues of the crossproduct matrix

$$\mathsf{C} = \begin{bmatrix} \int y_0^2(t)dt & \int y_0(t)y_i(t)dt \\ \int y_0(t)y_i(t)dt & \int y_i^2(t)dt \end{bmatrix} \tag{4}$$

is going to be nonzero. This suggests that we might choose warping function $h(t)$ as to minimize the smallest eigenvalue of the crossproduct matrix i.e.

$$F(h) = \mu_2(\mathsf{C}), \tag{5}$$

where $F(h)$ is continuous curve registration criteria casted into functional form and $\mu_2(\mathsf{C})$ is the smallest eigenvalue

of crossproduct matrix C. Thus for the general case represented by (3), minimization of smallest eigenvalue of the crossproduct matrix will lead us to ideal solution of completely registered curves and provides us the time warp function h.

The warping functions h are required by most applications to be both monotone and smooth. If in addition to being strictly increasing, we assume that h has an integrable second derivative, then h can be described by the homogeneous linear differential equation

$$\frac{d^2h}{dt^2} = w\frac{dh}{dt} \tag{6}$$

because a strictly monotone function has a nonzero derivative, and hence weight function w is simply $\frac{d^2h}{dt^2}/\frac{dh}{dt}$, or the relative curvature of h. This equation, subject to the requirement that $h(0) = 0$ and $h(T_0) = T_i$, has the solution

$$h(t) = C_1 \int_0^t \exp\left[\int_0^u w(v)dv\right] du \tag{7}$$

Note that constant C_1 is necessarily $T_i/[\int \exp \int w(T_0)]$.

In practice, we impose a penalty on the roughness of $w(t)$, or, equivalently, on $\int w(t)$ in our criteria $F(h)$. This is achieved by minimizing

$$F_\lambda(h) = F(h) + \lambda \int \left[\frac{d^m w(t)}{dt^m}\right]^2 dt \tag{8}$$

where $F_\lambda(h)$ is the composite minimization criteria consisting of minimization of smallest eigenvalue of crossproduct matrix $F(h)$ and the second term controls the smoothness of h. If $m = 0$, larger values of smoothing parameter λ shrink the relative curvature $w = \frac{d^2h}{dt}/\frac{dh}{dt}$ to zero, and therefore shrink $h(t)$ to t. Since the relative curvature measure w is scale free, appropriate values of λ tend not to vary much from one application to another. However, if we need to estimate derivatives of $h(t)$, it is better to work with higher values of m as chain rule will imply that we take the corresponding derivatives of $h(t)$ also. Specifically, if the first derivative is needed, using $m = 1$ will effectively penalize the total curvature, and thus keep it as smooth as desired. In the limit $\lambda \to \infty$, this will ensure that $w(t)$ is a constant. In practice, we use $m = 2$ as we want reliable estimate of higher order derivatives of h.

Figure 6 shows the results both before and after registration.

4 Conditions for Applicability

In §3 we showed how to estimate the time warp function between two or more events using curve registration on trajectories $\mathbf{y}(t)$. Now we describe the conditions required for curve registration to be physically meaningful. It is well known that trajectory of a moving point as seen in image sequences depends on camera viewpoint and its parameters. Even if camera parameters and viewpoint is fixed, point's trajectory as perceived in image sequences depend on its velocity.

A 3D point at any time t is represented by a 3-vector $\mathbf{X}(t) = [X(t), Y(t), Z(t)]^\mathsf{T}$. Its corresponding retinal point is represented by 2-vector $\mathbf{x}(t) = [x(t), y(t)]^\mathsf{T}$ or

$$x = f\tfrac{X}{Z}, \qquad y = f\tfrac{Y}{Z} \tag{9}$$

where f is the focal length and we have dropped t for simplicity of notation. Given that the focal length f is constant, the derivative of (9) w.r.t. time t is

$$\dot{y} = f\frac{\dot{Y}}{Z} - f\frac{Y}{Z^2}\dot{Z}, \tag{10}$$

where $\dot{y}$ denotes the derivative of y w.r.t. time t and so on. Similar results will hold for x-coordinate also. From (10), we see that $\dot{y}$ involves both $\dot{Y}$ and $\dot{Z}$ function. And hence supports above statement that a point trajectory $\mathbf{x}(t)$ in image depends upon its velocity even if camera parameters are fixed. To perceive it easily, consider $\dot{Y} = 0$ at time t_m then using (10)

$$\dot{y}|_{\dot{Y}=0} = -f\frac{Y}{Z^2}\dot{Z}$$

In other words, if at time t_m we have a minimum or maximum in Y, it does not imply that we have a minimum or maximum in image in general.

Thus, in general, we can not apply curve registration as trajectories in image plane depend on large number of variables (like camera viewpoint, velocity, etc.) and we can not control all of them. This is one more reason why we are opposed to using a static manifold in §2.3 for obtaining image plane trajectories $\mathbf{y}(t)$.

4.1 When Does it Work?

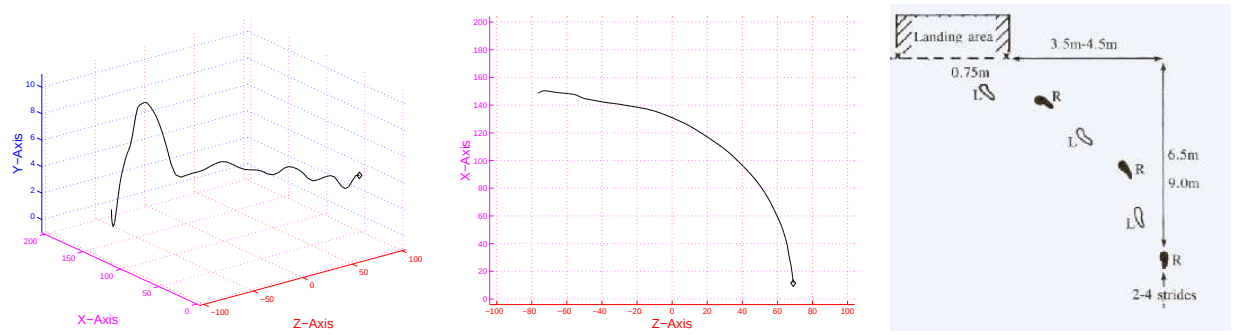
For curve registration to work for all possible cases i.e. changing camera parameters and different viewpoints, we would like that $y \propto Y$ and $x \propto X$ (or $\dot{y} \propto \dot{Y}$ and $\dot{x} \propto \dot{X}$) at all time instances t. We will now consider only y-coordinate. Similar results will hold for x-coordinate. The way $\dot{y}$ is presented in (10), $\dot{y} \propto \dot{Y}$ is mathematically impossible in general. However, if we allow focal length f to vary, the derivative of (9) w.r.t. time t is

$$\dot{y} = f\frac{\dot{Y}}{Z} + f\frac{Y}{Z}\left(\frac{\dot{f}}{f} - \frac{\dot{Z}}{Z}\right) \tag{11}$$

If $\dot{y} \propto \dot{Y}$ for all time instances t, we must have

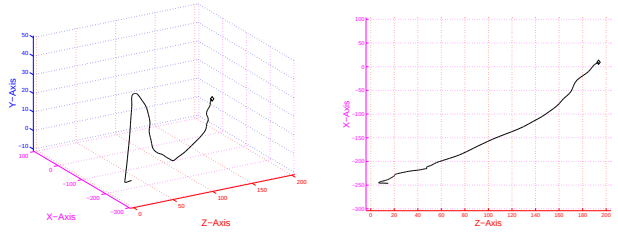
$$\frac{\dot{f}}{f} - \frac{\dot{Z}}{Z} = 0 \text{ or } \boxed{f \propto Z} \tag{12}$$

Thus, curve registration techniques will work if the camera zoom varies in such a way that focal length f at any time is proportional to Z-coordinate of the moving point at that time. In fact, it can be easily proved for general camera



(a) (b) (c)

Figure 2. 3D trajectory obtained for high jump athlete: (a) trajectory in 3D; (b) top view of trajectory; and (c) Top view of the optimal high jump trajectory. Compare (b) and (c).



(a) (b)

Figure 3. 3D trajectory obtained for pole vault athlete: (a) trajectory in 3D; (b) top view of the trajectory. Note that camera parameters does effect recovered 3D trajectories. Some problems at start are due to initial incorrect estimation of homographies.

motion (i.e. zooming, rotating and translating camera) that $\dot{y} \propto \dot{Y}$ and $\dot{x} \propto \dot{X}$ if and only if $f \propto Z$ at any time instant.

Another physical interpretation of $f \propto Z$ is that the object of interest must remain of *equal size* in all images. Thus, if $f \propto Z$, one can perceive true 3D trajectory as spanned by an object from just one camera. Note that this is a sufficient and necessary condition for trajectories in image plane to be independent of camera viewpoint and motion.

Most professionally produced sport videos keep objects of interest of the same size for better visualization. This is particularly true when the trajectory of the athletes are known in advance up to a good accuracy. In order to verify if condition (12) actually holds in practice, we solved the resulting ordinary differential equations numerically (see Appendix A) for pant-tilt and zooming cameras. The result is the trajectory spanned by the athlete's torso in 3D space from a single camera. Figure 2 shows the trajectory of a high jump athlete. As can be seen, the trajectory obtained with this method is very close to the optimal trajectory of the high jump figure 2-(c). Figure 3 shows the trajectory for a pole vault.

5 Automatic Indexing of Videos

The representation of human body motion as a time-varying curve together with the registration of two such curves extracted from two different video sequences allows us to automatically index key positions in the two videos.

For each video sequence, athlete's torso trajectory is extracted. Given two such curves, say $\mathbf{y}_1(t)$ and $\mathbf{y}_2(t)$, we compute the time warp function $h(t)$ such that $\mathbf{y}_1[h(t)] = \mathbf{y}_2(t)$ as described in §3. In practice, we apply the curve registration technique to velocities, i.e., $\mathbf{v}(t)$. One reason, as noted in §3, is as higher order derivatives vary much more than the function itself, it makes the registration process more robust. Second order derivatives can also be used. But it should be noted that imperfect tracking can create problems in estimating reliable higher order derivatives. Also by using velocity vectors $\mathbf{v}(t)$ we avoid explicit normalization and translation of the curves.

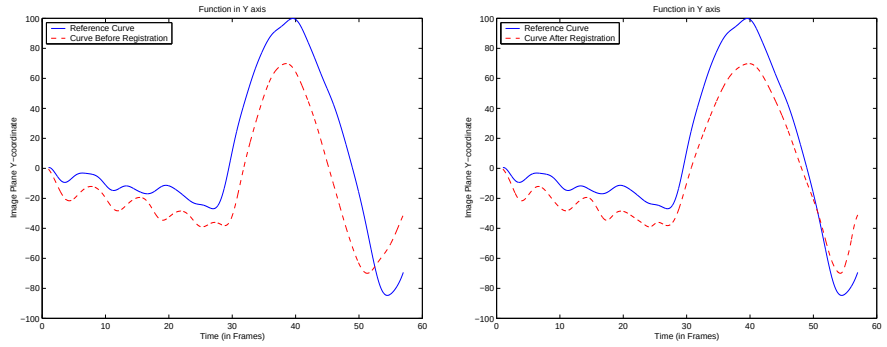
Key frames are manually selected in the first video. Using the estimated time warp function, time values corresponding to key frames as specified in the first video sequence are transformed into time values in the second video sequence.

Figure 4 shows two high jump video sequences where initial frames have been synchronized. Frames from the first video – (a) – are selected as key frames. Since the two curves corresponding to the respective trajectories are not registered, corresponding frames in the second video – (b) – do not correspond to the same key frames. Figure 6-(a) reflects this situation.

Figure 5 shows the results of indexing the two videos once curve registration has been applied. Manually labelled key postures in the first video are automatically indexed in the second video. Figure 6-(b) reflects this situation. In second video, we lost the track of the athlete for a while toward the end of the sequence. But as registration depends mostly on key landmark points along trajectories, it still manages to find correct key postures in the second video.

Figure 4. Few equally spaced frames of two high jump input video sequences. Note that viewpoint in second video sequence is quite different. Though starting point is same for two video sequences, but timings for jump are different.

Figure 5. Results of automatic key frames selection: (a) Key frames are selected manually in one video sequence; (b) Corresponding key frames obtained by applying curve registration to estimate time warp function.



(a) (b)

Figure 6. Curve registration results on two high jump trajectories. y coordinate of egomotion subtracted adaptive manifold based trajectory $\boldsymbol{y}(t)$ plotted against time t: (a): Two curves before registration. Solid line represents reference curve. Curve to register is represented as dashed line. This case corresponds to input video sequences as shown in fi gure 4; (b): Two curves after registration. Curve after registration represented as dashed line. Note that after registration features of dashed line curve (for e.g. maximums and minimums) align perfectly w.r.t. reference curve. Figure 5 shows the corresponding video sequences registered in time w.r.t. each other.

This methodology can be easily extended to tracking or indexing any body parts, joints, hand gestures and so forth.

6 Conclusions

In this paper, we presented a framework targeted toward human activity analysis when only one camera view is available. By analysis we mean possibilities: like indexing key positions in video sequences, recognition of events, performance analysis, comparison of two or more events, etc. In particular, we extracted spatio-temporal signatures associated with various events. Curve registration techniques are then applied to these spatio-temporal signatures to compute time warp functions between events. Registration of

curves to estimate time warp function though known to vision community, but has not been fully explored. One main reason is its dependency on camera parameters and viewing angle. In this paper we provide the necessary and sufficient conditions for viewpoint independence. Also, we performed experiments to check validity of these conditions in real life video sequences.

Next, we showed one possible application of the framework: automatic indexing of key positions in video sequences. Other possible applications can be: recognition of events, human body tracking, full 3D reconstruction of complex human motions and trajectories, etc. Also we believe that current framework can handle more subtle issues like performance analysis, comparison of two or more events, etc. more elegantly. The power of the approach results from the fact that it allows physically meaningful comparison of events at comparable time instances. In the future we intend to show, both systematically and experimentally, above promised possibilities.

References

[1] A. Bartoli, N. Dalal, B. Bose, and R. Horaud. From video sequences to motion panorama. In *Proceedings of the* IEEE *workshop on Motion and Video Computing, Orlando, Florida, USA*, 2002.

[2] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA*, pages 382–388, 1995.

[3] T. Darrell and A. Pentland. Space-time gestures. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, New York, USA*, pages 335–340, 1993.

[4] L. de Agapito, R. Hartley, and E. Hayman. Linear self-calibration of a rotating and zooming camera. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA*, 1999.

[5] R. I. Hartley. Self-calibration of stationary cameras. *International Journal of Computer Vision*, 22(1):5–23, 1997.

[6] M. Isard and A. Blake. Condensation-conditional density propagation for visual tracking. *IJCV*, 29:5–28, 1998.

[7] C. Myers, L. Rabinier, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE *Transactions on Acoustic, Speech and Signal Processing*, 28(6):623–635, 1980.

[8] J. O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society B*, 60:351–363, 1998.

[9] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer-Verlag, 2002.

[10] S. Seitz and C. R. Dyer. Affi ne invariant detection of periodic motion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA*, pages 970–975, 1994.

[11] H.-Y. Shum and R. Szeliski. Systems and experiment paper: Construction of panoramic mosaics with global and local alignment. *International Journal of Computer Vision*, 36(2):101–130, February 2000.

[12] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Proceedings of the* IEEE *International Symposium on Computer Vision, Coral Gables, Florida, USA*, pages 265–270, 1995.

A 3D Trajectory Generation

We now show how to estimate 3D trajectory of a point given $f(t) \propto Z(t)$ at any time. We assume that 3D coordinate reference frame is attached to camera such that image plane is normal to Z-axis. Also we assume we have pan-tilt and zoom camera and pan, tilt angles and focal length values at any time t are known (see §2.1). However note that results presented below can be very easily extended to general camera motion.

A 3D point at any time is represented by $\mathbf{X}^t = [X^t, Y^t, Z^t]^\mathsf{T}$. Here superscript denotes point coordinates as defined in instantaneous camera reference frame at time t. It can be proved easily for general camera motion that if $f \propto Z^t$ then $\dot{y} \propto \dot{Y}^t$ and $\dot{x} \propto \dot{X}^t$. Thus we can write

$$\lambda \begin{bmatrix} x \\ y \\ f \end{bmatrix} = \begin{bmatrix} X^t + t_x \\ Y^t + t_y \\ Z^t \end{bmatrix} \text{ or } \lambda\xi = \mathbf{Y}^t, \qquad (13)$$

where $\xi = [x, y, f]^\mathsf{T}$, $\mathbf{Y}^t = [X^t + t_x, Y^t + t_y, Z^t]^\mathsf{T}$, λ is a constant unknown scale factor, t_x and t_y are constant unknown translational factors. This equation is same as standard perspective camera projection expression in homogeneous coordinates. However, note that $[x, y]^\mathsf{T} = \mathbf{y}(t)$ are coordinates of feature point curve as described in §2.3 i.e. egomotion subtracted point trajectory as seen in the image plane.

For the case of a rotating camera, from (13), we have

$$\lambda\xi = \mathsf{R}\mathbf{Y}^{t-\Delta t}, \qquad (14)$$

where R is the camera rotation in time Δt. As $\Delta t \to 0$,

$$\mathsf{R} \to \mathcal{I}_{3\times 3}, \qquad (15)$$

where $\mathcal{I}$ is identity matrix. Taking derivative of (14) w.r.t. time t and using (15)

$$\lambda\dot{\xi} = \dot{\mathbf{Y}} + \Omega \times \mathbf{Y}, \qquad (16)$$

where we have dropped superscript t, Ω is 3-vector resprenting angular velocity and $\times$ denotes the vector cross-product. We solve the above ordinary differential equation (16) numerically (up to a constant unknown scale factor λ and unknown initial conditions t_x and t_y) to obtain trajectory of a point in 3D given rotation matrix and focal length. Figure 2 and figure 3 shows the results obtained.